DOCUMENT RESUME

ED 473 530 TM 034 738

AUTHOR	van der Linden, Wim J.
TITLE	Estimating Equating Error in Observed-Score Equating. Research Report.
INSTITUTION	Twente Univ., Enschede (Netherlands). Faculty of Educational Science and Technology.
SPONS AGENCY	Law School Admission Council, Princeton, NJ.
REPORT NO	RR-0Z-03
PUB DATE	2002-00-00
NOTE	40p.
AVAILABLE FROM	Faculty of Educational Science and Technology, University of Twente, TO/ OMD, P.O. Box 217, 7500 AE Enschede, The Netherlands. E-mail: Fox@edte.utwente.nl.
PUB TYPE	Reports - Research (143)
EDRS PRICE	EDRS Price MF01/PC02 Plus Postage.
DESCRIPTORS	College Entrance Examinations; *Equated Scores; *Error of Measurement; Estimation (Mathematics); *Item Response Theory; Law Schools
IDENTIFIERS	Conditional Observed Scores; *Equipercentile Equating; *Law School Admission Test

ABSTRACT

Traditionally, error in equating observed scores on two versions of a test is defined as the difference between the transformations that equate the quantiles of their distributions in the sample and in the population of examinees. This definition underlies, for example, the well-known approximation to the standard error of equating by Lord (1982). However, it is argued that if the goal of equating is to adjust the scores of examinees on one version of the test to make them indistinguishable from those on another, equating error should be defined as the degree to which the equated scores realize this goal. Two equivalent definitions of equating error based on this criterion are formulated. These definitions can be used to evaluated existing equating methods and derive new methods if the response data fit an item-response theory model. An evaluation of the traditional equipercentile equating method and two new conditional methods for tests from a previous item pool of the Law School Admission Test showed that, under a variety of conditions, the equipercentile method tends to result in a serious bias in the equated scores, while the new methods are practically free of any bias. (Contains 5 figures and 14 references.) (Author/SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 473 530

TM TMTR 0080

Estimating Equating Error in Observed-Score Equating

Research Report 02-03

Wim J. van der Linden

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

J. Nelissen

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

TM034738

faculty of
EDUCATIONAL SCIENCE AND TECHNOLOGY

University of Twente

Department of Educational Measurement and Data Analysis

BEST COPY AVAILABLE

Estimating Equating Error in Observed-Score Equating

Wim J. van der Linden

This study received funding from the Law School Admissions Council (LSAC). The opinions and conclusions contained in this paper are those of the other and do not necessarily reflect the policy and position of LSAC. The paper was completed while the autor was a Fellow at the Center of Advanced Study in the Behavioral Sciences, Stanford, CA. The author is indebted to the Spencer Foundation for a grant awarded to the Center sepport his Fellowship. The computational assistance of Wim. M.M. Tielen is gratefully acknowledged.
Requests for reprints should be sent to W.J. van der Linden, Department of Educational Measurements and Data Analysis, University of Twente, P.O. Box 217, 7500 AE Enschede, THE NETHERLANDS. Email: w.j.vanderlingen@edte.utwente.nl

Abstract

Traditionally, error in equating observed scores on two versions of a test is defined as the difference between the transformations that equate the quantiles of their distributions in the sample and in the population of examinees. This definition underlies, for example, the well-known approximation to the standard error of equating by Lord (1982) . However, it is argued that if the goal of equating is to adjust the scores of examinees on one version of the test to make them indistinguishable from those on another, equating error should be defined as the degree to which the equated scores realize this goal. Two equivalent definitions of equating error based on this criterion are formulated. These definitions can be used to evaluate existing equating methods and derive new methods if the response data fit an item-response theory model. An evaluation of the traditional equipercentile equating method and two new conditional methods for tests from a previous item pool of the Law School Admission Test (LSAT) showed that, under a variety of conditions, the equipercentile method tends to result in a serious bias in the equated scores, whereas the new methods are practically free of any bias.

Key words: classical test theory (CTT); computerized adaptive testing; conditional equating; equating error; item response theory (IRT); marginal equating; observed-score equating.

Estimating Equating Error in Observed-Score Equating

The goal of observed-score equating is to adjust the observed number-correct scores on a new version of a test to make them indistinguishable from the scores the examinees would have had if they had taken an old version. Key questions in observed-score equating are thus: When are scores of examinees on two versions of a test indistinguishable from each other? And what score transformation does realize this goal best?

In the practice of observed-score equating, the benchmark among the methods of observed-score equating is generally considered to be equipercentile equating with a random-groups design. This method requires the distribution functions of the equated scores and the scores on the old version of the test to be identical for the population of examinees. To estimate the transformation, the two versions are administered to different random samples from the population and the transformation is inferred from the sample distributions.

More formally, equipercentile equating can be defined as follows. Let Y be the observed score on the new version of a test for a random examinee from a population that has to be equated to the observed score X on an old version. These variables take possible values y and x, respectively. For convenience, we will also use Y and X to denote the versions of the tests themselves. The distribution functions for the two variables are denoted as $F_Y(y)$ and $F_X(x)$. For simplicity, throughout this paper we will assume that both functions are continuous and monotone. The transformation to be estimated equates the quantiles in the two distributions. If the distribution functions are monotonically increasing, the transformation is

$$x = \varphi(y) = F_X^{-1}(F_Y(y)), \tag{1}$$

where $\varphi(y)$ is the y score equated to the scale of X (Braun & Holland, 1982). The transformation in (1), which is monotonically increasing, maps the yth quantile in a

distribution with distribution function $F_Y(.)$ to the same quantile in a distribution with distribution function $F_X(.)$.

Note for future reference that the transformation in (1) is from scale y to scale x. However, if it is applied to y, the random variable Y is transformed into a random variable $\varphi(Y)$ with a population distribution identical to the distribution of X.

Because the distribution functions of X and Y have to be estimated, the actual transformation used to equate the scores of X and Y is

$$x = \widehat{\varphi}(y) = \widehat{F}_X^{-1}(\widehat{F}_Y(y)), \tag{2}$$

where $\widehat{F}_Y(y)$ and $\widehat{F}_X(x)$ are estimates of the two population distribution functions.

One option to estimate the population distribution functions in (1) is to use their sample equivalents. However, for a typical test length, this method involves the estimation of a large number of parameters. (To be precise, n parameters for the distribution function of a test of n items, namely the population proportions for each of the number-correct scores 0,1,...,n minus one because their sum is constrained to be equal to one.) Thus, to get stable sample distributions functions, large samples are required, particularly if the tails of the distributions, where the proportions of examinees in the population tend to be smaller, matter. The sample size can be reduced somewhat by using a sensible smoothing technique, but the danger of exchanging inaccurate for biased estimator is inherent in any such technique. An alternative option is to assume that the population distribution functions belong to a parametric family of functions and use the sample to estimate their parameter values. If statistical techniques are available to test the assumption against response data, this option becomes efficient. This option is used in the empirical examples later in this paper.

The estimator in (2) implies the following definition of equating error:

$$\varepsilon(y) = \widehat{\varphi}(y) - \varphi(y) = \widehat{F}_X^{-1}(\widehat{F}_Y(y)) - F_X^{-1}(F_Y(y)), \tag{3}$$

that is, all equating error is due to sampling fluctuations in the estimators $\widehat{F}_F(y)$ and $\widehat{F}_X(x)$. If these estimators converge to their population equivalents, which they do if the sample distribution function is used as the estimator of the population function, equating error vanishes and the equating becomes perfect. Observe that in (3) error is random across sampling of examinees from the population. Also, error is a function of y; for different values of y equating error takes different values.

The definition in (3) is indeed the definition underlying the literature on observed-score equating error, which was put on statistical footing in Lord's (1982) seminal paper, in which he presented a large-sample approximation to the standard error of equating. The approximation is given by:

$$Var(\widehat{\varphi}(y)) \approx \frac{F_Y(y)[1 - F_Y(y)]}{f_X(x')} \left(\frac{1}{N_X} + \frac{1}{N_Y} \right), \tag{4}$$

where

$$x' = F_X^{-1}(F_Y(y))$$

and N_X and N_Y are the sizes of the samples from the distributions of X and Y. Lord's derivation of (4), which assumes that the distribution functions in (2) are estimated by their sample equivalents, is based on a variance decomposition for the left-hand side of (4) to deal with the fact that (2) is a composite function with random error in each component. The result in (4) is for continuous test scores; for an approximation for discrete scores, see Lord (1982).

Basically the same definition of equating error is found in Kolen and Brennan (1995, p. 212), albeit these authors arrive at the definition from the opposite direction. Given any transformation $\varphi(.)$, they define its application to a sample of examinees on test Y as the equated score $\widehat{\varphi}(y)$. They then go on and, in a step reminiscent of the definition of an examinee's true score in classical test theory, define the true equated score as the expected value of the equated score across sampling, $\mathcal{E}[\widehat{\varphi}(y)]$. It follows that the equating

error associated with y is equal to

$$\varepsilon(y) = \widehat{\varphi}(y) - \mathcal{E}[\widehat{\varphi}(y)] \tag{5}$$

for all values of y. This definition is identical to the one in (3), provided $\widehat{\varphi}(y) = \varphi(y)$, that is $\widehat{\varphi}(y)$ is an unbiased estimate of $\varphi(y)$ for all values of y, which holds if the distribution functions in (1) are estimated by their sample equivalents. To estimate the small-sample standard error associated with (5), Kolen and Brennan recommend using a (parametric) bootstrap estimator.

Formulation of Problem

The above definitions of equating error are puzzling for the following reasons. First, if the goal of observed-score equating is to yield equated scores on the new version of the test that, for all examinees, are indistinguishable from the scores on the old version, the definition of error should be based on this goal. That is, error should be defined as a measure of the differences between the scores of examinees on the two version of the test.

Second, the definitions of equating error in (3) and (5) are based on a necessary condition but not on a necessary and sufficient condition for successful equating. If there are no equating errors in the sense that the equated scores on the new version of the test and observed scores on the old version are indistinguishable for each examinee in a population, the population distributions of these two scores are identical. However, the reverse does not hold: If for a population the two distributions are identical, it is not necessary that the equated scores are indistinguishable from the scores on the old version of the test for each examinee.

Third, the approach underlying (5) seems even to be circular. It accepts any function of y as an equating transformation and defines its expectation across sampling as the true equated score against which the sample result should be evaluated. This approach works fine in classical test theory where, for lack of an external criterion, we replace the observed score of an examinee by its expectation as the parameter of interest, but not in observed-

score equating, where we do have an external criterion in the form of the observed score on the old version of the test. Also, (5) implies that equated scores can never be biased, even if they result from an obviously wrong transformation, such as $\varphi(y) = c$, where c is the same arbitrary number for all values of y. In fact, for $\varphi(y) = c$ (5) even implies error-free equating for all values of y!

The problem addressed in this research is how to define equating error if the criterion of successful equating is that one should not be able to distinguish between the equated scores on the new version of the test and the scores on the old version. Two equivalent definitions of equating error based on this criterion are formulated. Also, it is shown that if these definitions are embedded in the framework of item response theory (IRT), it becomes possible to evaluate equating error for any type of equating transformation. The procedure is illustrated for the transformation in (2) and two alternative equating transformations that are introduced below using tests that were systematically varied in some of their properties.

Definition of Equating Error

It is now made more precise what is meant by observed scores on two versions of a test being "indistinguishable". Let $\mathcal{P}$ be the population of examinees from which we sample and p an arbitrary examinee in this population. A basic assumption in test theory is that for each examinee p the observed scores X_p and Y_p are random variables, that is, outcomes of test administrations that show random variation over (hypothetical) replications. These random variables are denoted as X_p and Y_p. The transformation $\varphi(y)$ thus yields a new random variable, $\varphi(Y_p)$, which is the equated score for examinee p. Observe that these random variables are different from the variables used to define the equipercentile transformation in (1).

The following definition of "indistinguishable" follows immediately the assumption that X_p and Y_p are random: Equated scores on the new version of the test and scores on

the old version are indistinguishable from each other for a population of examinees $\mathcal{P}$ if

$$\varphi(Y_p) \text{ and } X_p \text{ are identically distributed for all } p \in \mathcal{P}. \tag{6}$$

This definition was documented earlier as a criterion of equating by Lord (1980), who called it the equity criterion and claimed that no transformation satisfying this criterion exists, unless the two test versions are parallel and equating is not needed. However, as will be explained later in this paper, this claim was based on an unnecessary assumption about the nature of the transformation.

The following two sections present definitions of equating error that follow directly from (6).

Error in Equated Scores

The requirement in (6) implies that the distribution functions of $\varphi(Y_p)$ and X_p be identical. A natural definition of error in an equated score is the difference between these functions, $F_{\varphi(Y_p)}(\varphi(y)) - F_{X_p}(x)$, where, for each value of y, the distribution function $F_{X_p}(x)$ is evaluated at the value of x to which y is equated. More compactly, the definition of error in equated score is thus

$$\varepsilon_{p1}(y) = F_{\varphi(Y_p)}(\varphi(y)) - F_{X_p}(\varphi(y)). \tag{7}$$

Note that, contrary to what might be expected intuitively, this definition of error does not lead to a not to a single number but to function of y for each examinee. The reason is the random nature of the scores X_p and Y_p. Any attempt to further reduce this error function may lead to loss of important information. For example, if we focused only on the maximum value of the $\varepsilon_{p1}(y)$ over y or replaced it by a measure of the area between the two distribution functions in (7), we would never know for what part of scale y the equating transformation has the potential to distort the equated scores.

Error in Equating Transformations

If the condition $\varepsilon_{p1}(y) = 0$ is imposed on (7) for all y, and the equality is solved for $\varphi(y)$, the solution is an equating transformation from Y to X which, according to the definition in (7), is free of error.

The condition $\varepsilon_{p1}(y) = 0$ gives

$$F_{\varphi(Y_p)}(\varphi(y)) = F_{X_p}(x)$$

or

$$x = F^{-1}_{X_p}(F_{\varphi(Y_p)}(\varphi(y)).$$

Because

$$F_{\varphi(Y_p)}(\varphi(y)) = F_{Y_p}(y),$$

it thus holds that

$$x = \varphi^*_p(y) = F^{-1}_{X_p}(F_{Y_p}(y)). \tag{8}$$

Observe that this transformation has the same shape as the equipercentile transformation in (1). This fact should not come as a surprise. The equipercentile transformation, which more appropriately should be called the quantile or the Q-Q transformation (Wilks & Gnanadesikan, 1968), can be used to transform any (continuous and monotone) distribution function into any other. However, the fundamental difference between (1) and (8) is that the former is applied to the score distributions of a population of examinees and the latter to the distributions of the observed scores of a single examinee.

The equating transformation $\varphi^*_p(y)$ in (8), which for obvious reasons is referred to as the true equating transformation in the remainder of this paper, can be used to evaluate the error in any other transformation. Let $\varphi(y)$ be a proposed equating transformation.

As an alternative to (7), error in $\varphi(y)$ can be defined as

$$\varepsilon_{p2}(y) = \varphi(y) - F^{-1}_{X_p}(F_{Y_p}(y)). \tag{9}$$

The two definitions of error in (7) and (9) are equivalent: $\varepsilon_{p2}(y)$ shows for which part of the scale equating transformation $\varphi(y)$ goes wrong, for instance, overstretches the scale of Y, where $\varepsilon_{p1}(y)$ shows the mismatch in the distribution of the equated score $\varphi(Y_p)$ that is the result. Observe that, like (7), the definition in (9) entails a function of y for each examinee. However, both functions have a different range: $\varepsilon_{p1}(y)$ takes values in [0,1], whereas $\varepsilon_{p2}(y)$ takes values on the scale of X. In the empirical examples later in this paper, because of the possibility to interpret equating errors directly on the scale of Y, $\varepsilon_{p2}(y)$ was used to evaluate different equating transformations under a variety of conditions.

Discussion

It is important to realize that this definition of indistinguishable scores in (6) requires only that $\varphi(Y_p)$ and X_p be identically distributed. As an alternative, one might feel inclined to impose the stronger requirement that $\varphi(Y_p)$ and X_p be *identical*, that is,

$$\varphi(Y_p) = X_p \text{ for all } p \in \mathcal{P}. \tag{10}$$

If this requirement were to hold, equating error would have to be defined as

$$\mathcal{E}_p = \varphi(Y_p) - X_p. \tag{11}$$

However, $\varphi(Y_p)$ and X_p are never identical. The well-known property of conditional (or local) independence of test scores even implies that they are independent. Because of this property, the definition in (11) leads to a standard error of equating equal to

$$[Var(\varphi(Y_p) + Var(X_p)]^{-1/2}, \text{ for all } p \in \mathcal{P}. \tag{12}$$

This expression is minimal for the transformation $\varphi(y) = c$, with a minimum equal to $[Var(X_p)]^{-1/2}$, which is the standard error of measurement for X_p. This implication shows that (11) can not be used as a meaningful definition of equating error.

Another alternative to the error definitions in (7) and (9) might seem to define equating error as the difference between the equated score associated with the examinee's *realized* score $Y_p = y$ on test Y and his/her score on X. This choice would amount to the conception of equating error as a conditional random variable given $Y_p = y$,

$$\mathcal{E}_p \mid y = \varphi(y) - X_p. \tag{13}$$

However, as $\varphi(y)$ is now a constant, (13) implies the same distribution of equating error for each possible value y, no matter the equating transformation. In this case, the minimum standard error of equating for the previous alternative definition, $[Var(X_p)]^{-1/2}$, would thus hold for any transformation. This implication shows that (13) can not be used as a meaningful definition of equating error either.

Estimating Equating Error

So far, the results have only been theoretical. In practice, the distributions of the observed scores X_p and Y_p are unknown, and the only datum available for each examinee in a random-groups design is one realization of the scores $X_p = x$ or $Y_p = y$. Without any further assumptions, it is thus impossible to estimate equating error. We will now look into assumptions that do allow us to do so.

Classical Observed-Score Equating

A natural approach to getting more data about the distributions of observed scores of examinees is to pool scores across examinees. In the framework of classical test theory (CTT), it might seem attractive to pool the data of examinees with the same true score on test X and Y. Let τ_{X_p} and τ_{Y_p} be the true scores of examinee p on X and Y, respectively. In CTT, these true scores are defined as the expected observed scores of the

examinee on these tests. Pooling examinees with the same true scores amounts to forming subpopulations in $\mathcal{P}$ with examinees $p \in \mathcal{P}$ for which $\tau_{X_p} = \tau_X$ or $\tau_{Y_p} = \tau_Y$ for some values for τ_X and τ_{Y_p}. Technically, the approach means that the distribution functions in (7) and (9) are no longer indexed by p but are replaced by the conditional distribution functions $F_{X|\tau_X}$ and $F_{Y|\tau_Y}$.

However, an implementation of this approach would have to deal with several obstacles. For example, it is not known how to test the assumption of identical conditional observed score distributions for all examinees with the same true score on which the approach rests. In addition, given the fact that only one observed score is available for each examinee, it seems impossible to find reasonable estimators for the true scores of the examinees on the two test versions. Finally, to calculate the errors in (7) or (9) it must be known how to pair the true score on Y to the one on X, but it seems impossible to infer this relation from the available data.

IRT Observed-Score Equating

A more practical alternative to implementing the idea of pooling scores is offered by item response theory (IRT). The models in IRT are based on stronger assumptions about the response to the items in X and Y than CTT, but powerful statistical tests exists to check them. In the section with the empirical examples below, the 3-parameter IRT model is assumed to hold for both versions of the test simultaneously:

$$p_i(\theta) = \Pr(U_i = 1 \mid \theta) = c_i + (1 - c_i)\{1 + \exp[-a_j(\theta - b_j)]\}^{-1}, \qquad (14)$$

where U_i is a binary variable for the response of the examinee to item i, $\theta \in (-\infty, \infty)$ represent the examinee's ability level, and $a_i \in [0, \infty)$, $b_i \in (-\infty, \infty)$, and $c_i \in [0, 1]$ are the discriminating power, difficulty, and guessing parameter for item i (Lord, 1980).

Under the model in (14), the distribution functions in (7) and (9) are replaced by the functions of Y and X given θ, which will be denoted as $F_{Y|\theta}(y)$ and $F_{X|\theta}(x)$, respectively. Because the model is assumed to hold for X and Y simultaneously, the

conditioning variable θ is common. As a result, unlike CTT, there is no need to infer a functional relation between the examinee parameters for both test versions; given θ, it is automatically clear what distribution of X is associated with what distribution of Y.

For IRT models for dichotomously scored responses on the items, such as the one in (14), the distributions of Y and X given θ belong to the generalized binomial family (also known as the compound binomial family; e.g., Lord, 1980). This family does not have distribution functions that can be expressed in closed form, but its members can easily be calculated using a recursive procedure in Lord and Wingersky (1984). The procedure is based on the fact that the probabilities at $X = x$ are given by coefficient of factor t^x in the generating function

$$\prod_{i=1}^{n}[q_i(\theta) + tp_i(\theta)], \tag{15}$$

with $q_i(\theta) = 1 - p_i(\theta)$.

If this procedure has been used to calculate $F_{Y|\theta}(y)$ and $F_{X|\theta}(x)$ for a given value of θ, equating error in (7) and (9) can be calculated as

$$\varepsilon_1(y;\theta) = F_{\varphi(Y)|\theta}(\varphi(y)) - F_{X|\theta}(\varphi(y)) \tag{16}$$

and

$$\varepsilon_2(y;\theta) = \varphi(y) - F^{-1}_{X|\theta}(F_{Y|\theta}(y)), \tag{17}$$

respectively. Observe that (16)-(17) are now functions of y as well as θ.

Conditional Equating Methods

The error definition in (17) is based on the following set of true equating transformations

$$\varphi^*(y;\theta) = F^{-1}_{X|\theta}(F_{Y|\theta}(y)), \theta \in R. \tag{18}$$

This set was proposed directly for IRT observed-score equating in van der Linden (2000, Proposition 1), who proved that it meets all known criteria of perfect equating, namely (1) equity of equating for each examinee, (2) symmetry in X and Y, (3) population invariance, and (4) identical order of examinees on X and $\varphi(Y)$ (for the first three criteria, see Harris & Crouse, 1993, and Kolen & Brennan, 1995). This paper also presents graphical examples of the transformation for different tests as well as generalizations to tests that fit only multidimensional IRT models.

The fact that (18) meets the criterion of equity seems to contradict Lord's (1980) theorem, which claims that, except for the trivial case of two parallel test versions, such transformations do not exist. However, implicit in Lord's theorem is the assumption that the transformation should be the same function for all examinees, whereas (18) allows for different transformations for different examinees.

The critical feature of (18) relative to the equipercentile transformation in (8) is the conditioning on the examinee's ability θ. Equating methods with this feature will be called conditional equating methods, whereas the traditional equipercentile method in (1)-(2) will be referred to as a marginal equating method. The idea to condition an equating procedure on other variables than the observed test scores was already presented in Wright and Dorans (1993). A formal framework for doing so is presented in Liou, Cheng, and Li (2001). These authors motivated their approach by the wish to improve equating by accounting for relevant differences between examinees. Statistically, their idea amounts to the use of equating transformations conditional on values for background variables that describe relevant difference between examinees. In fact, the set of transformations in (18) takes this logic one important step further. Its transformations are conditional on the most relevant variable available: the ability of the examinees measured by the two versions of the test.

The error functions in (16)-(17) can be calculated for any given equating transformation as soon as the items in the two versions of the test have been calibrated using a sufficiently large sample of examinees. These functions are thus easily available

to evaluate a new equating procedure or to choose the best procedure for an equating study from an available set of candidates.

Although the set of true transformations in (18) is known as soon as the items have been calibrated, it is impossible in an actual equating study to pick the correct transformation from the set for a given examinee because his/her true value of θ is not known. Nevertheless, (18) is useful because it could suggest transformations that have smaller equating error than the estimated marginal equipercentile transformation in (2). In the next sections two suggestions by van der Linden (2000) are discussed. Each transformation deals in a different way with the fact that θ in (18) is unknown.

Estimated Conditional Equating

A simple approximation to (18) is to replace θ by an estimated inferred from the examinee's response vector. Let $\widehat{\theta}$ be such an estimate. In the empirical examples below, the expected a posteriori (EAP) estimator was used to estimate θ. A possible equating function, following upon the substitution of $\widehat{\theta}$ into (7), is thus:

$$\varphi(y;\widehat{\theta}) = F^{-1}_{X|\widehat{\theta}} F_{Y|\widehat{\theta}}(y),\ \widehat{\theta} \in R. \tag{19}$$

This transformation is based only on observable quantities. In an actual equating study, the following steps have to be taken to calculate an equated score for an examinee: (1) estimating the examinee's value of θ from his/her response vector; (2) calculating the true transformation in (18); and (3) using this transformation to calculate the equated score associated with the examinee's observed number-correct score $Y = y$.

Posterior Expected Conditional Equating

The transformation in (19) uses the mean of the posterior distribution of θ but ignores the remaining uncertainty about θ in this distribution. From a Bayesian perspective, it seems fair to acknowledge this uncertainty and take the expectation of (7) over the posterior distribution as an alternative to (19). If $f_{\Theta|u_1,\ldots,u_n}(\theta)$ denotes the density of the posterior distribution of θ for response vector $(u_1, \ldots, u_n)$, the equating transformation

becomes

$$\varphi(y; u_1, ..., u_n) = \int F^{-1}_{X|\theta} F_{Y|\theta}(y) f_{\Theta|u_1,...,u_n}(\theta) d\theta. \quad (20)$$

Discussion

Observe that the error functions in (19) and (20) are dependent on the response vector of the examinee. Because we are generally not interested in the evaluation of these functions for a single response by an examinee, it makes sense to evaluate them over random responses, for example, to examine their bias or average error across responses. This was done in the empirical examples below.

Statistically, the conditional transformations in (19)-(20), though never as good as the true transformation in (8), are expected to perform better than the marginal equation in (2). One reason is the bias in the marginal transformation due to its dependency on the population distribution of θ. Another is that the fact that the conditional transformations exploit the full information in the response vector of the examinee to find his/her he equated number-correct score, whereas the marginal transformation uses only the part of the information in the number-correct scores. Finally, we expect the conditional transformations to have better behavior for increasing test length. If the test length increases, the number of parameters that define the distribution functions on which the marginal transformation in (2) is based increases and less data per parameter becomes available to estimate them. On the other hand, if the test length increases, the point estimate of θ in (19) becomes more accurate and the posterior distribution of θ (20) degenerates at the true value of θ. Because, for each value of y, the equating transformations $\varphi^*(y; \theta)$ are continuous in θ, (19) and (20) converge to the true equations in (18).

Empirical Examples

Under a variety of conditions, response vectors for examinees on two different versions of a test were simulated and their number-correct scores were equated using

the marginal equating method in (1) as well as the estimated conditional and posterior expected conditional equating method in (19)-(20). The results were evaluated using the bias or average of the error defined in (17) across examinees at the same value of θ. Given the setup of the equating study, which is explained below, the marginal equating method had constant error for each value of θ, but for simplicity we will refer to this error also as bias.

The bias in the three transformations was evaluated for tests varying on the following factors:

(1) Length of X and Y;

(2) Difficulty of items in Y relative to those in X;

(3) Discriminating power of items in Y relative to those than in X;

(4) Size of error in parameter estimates of items in Y and X;

(5) Adaptive or fixed format for Y.

The first four factors were included in the study to identify possible aspects of tests critical with respect to the difference between results from marginal and conditional equating. When the effects of the first three factors were studied, all values of the item parameters were treated as if they were the true values. The fourth factor was included in the study to examine the effects of this assumption. The last factor was added because, even though examinees get different items if Y is adaptive, the conditional transformations in (18)-(19) put all examinees' number-correct scores on the same scale as X. This feature means that conditional equating is a potential alternative to the test characteristic curve method currently in use to report scores on adaptive tests as number-correct scores on a paper-and-pencil reference test or to equate score when examinees have the choice between an adaptive and a paper-and-pencil version of the same test (Lawrence & Feigenbaum, 1997; Segall, 1997; van der Linden, 2001).

Method

All versions of X and Y were derived from two blocks of 20 items selected from previous forms of the Law School Admission Test (LSAT). The 20-item, 40-item and 60-item tests in the conditions with varying test length consisted of one, two and three times

the same block. This setup guaranteed conditions with homogenous test lengthening and no confounding between test length and test composition.

The same 40-item versions of test X and Y were used as a standard for comparison in all conditions . The conditions with more and less difficult versions of the items in Y were simulated by subtracting and adding .5 to the values of parameter b_i for the items in the standard test, respectively. The conditions with more and less discriminating items in Y were simulated by multiplying their values for parameter a_i by .5 and 2.0, respectively. The condition with the smaller errors in the estimated value of the item parameters in X and Y was simulated by adding random numbers from [-.15, .15] to the values of the items in the two standard tests for parameters a_i and b_i, and from [-.10, .10] to the values for parameter c_i. The condition with larger error was simulated by taken the intervals twice as wide. In both conditions, the values of the estimates of a_i and c_i were set equal to .10 and .00 if the results was lower than these values.

The adaptive version of Y was a 40-item adaptive test simulated from a previous pool of 678 items from the LSAT. The ability estimator in the adaptive test was the EAP estimator with a uniform prior over [-4,4], which was always initiated at $\theta = 0$. The items were selected using the maximum-information criterion (van der Linden & Pashley, 2000).

Test administrations of X and Y were simulated for 5,000 examinees for each value θ=-2.00, -1.50, ..., 1.50, 2.00. The error functions in (17) were calculated for each examinee and the bias in the transformations was estimated as the averaged error in the functions over all examinees at the same value of θ. The number of simulations for each value of θ was large enough to get stable estimates of these bias functions.

The marginal equating transformation in this study was calculated generating the conditional observed-score distributions of X and Y given θ using (15) and then averaging these functions over the θ values (for another application of this method in equating, see Zeng & Kolen, 1995). The transformation was thus not the version estimated from the sample distribution functions in (2) but the population version in (1). The bias in the marginal equating transformation was therefore entirely due to the difference between

(1) and the true transformations in (18), and not to estimation error in the distribution functions in (2). As a consequence, the comparison between the results for the marginal and conditional methods is thus somewhat conservative in the sense that the marginal method was based on our knowledge of the true θ values of the examinees whereas for the conditional methods estimates of θ were used.

Results

The results for the different lengths of the test are displayed in Figure 1. The two conditional equating methods had already negligibly small bias at $n = 20$ for all values of θ, which further decreased with the length of the test. At $n = 20$, the bias in the posterior expected conditional equating method, though still less than one score point on X for all values of y, was noticeable larger than for the estimated conditional method, but the difference disappeared for the larger tests. The effect is believed to be due to the Bayesian nature of the former, which involves a larger bias as a price to be paid for a smaller accuracy.

[Figure 1 about here]

The plots for the marginal equating methods in all conditions yielded curves that were generally ordered in θ, with the curves for the higher values of θ more to the left. As expected, in Figure 1 these plots showed considerable bias at $n = 20$ for all values of θ, which further increased with the length of the test. Also, all curves showed a typical "wave", which is the result of the difference in shape between the marginal transformation in (1) and the true conditional transformations in (18).

Note that Figure 1 does not display functions over the entire range of y for all values of θ. The reason is the value of the conditional probability functions of X and Y given θ. For values smaller than .0001, it was decided that in a practical application the equating transformation would be unstable outside this range because too few examinees would be available and reporting of bias would no longer be relevant.

The difference in the values of the difficulty parameters for the two versions of the test resulted only in a shift in the wave of the curves for the marginal equating transformation.

This shift can be explained by the difference in shape between the marginal equating transformations for X less and more less difficulty than Y. In the former case, the transformation is concave; in the latter, it is convex.

[Figure 2 about here]

Interesting results were obtained for the conditions with a change in the values of the item discrimination parameter for the new test, Y, relative to the old test, X. Generally, lower values for this parameter means both wider observed number-correct distributions and less accurate estimators of θ. The first effect is visible in the plots for the marginal equating method in Figure 2, which shows curves over larger ranges of y values than in the previous plot. The second effect explains the larger errors in the conditional equating transformations, which for this case became substantial, particularly for the lower values of θ. For larger values of the discrimination parameter the opposite was observed: the marginal equating method still had large bias but the curves were defined over much smaller ranges of values of y whereas the conditional methods became virtually error free. However, it should be remembered that the changes in the values of the discrimination parameter in these two conditions relative to the values of 40-item test in Figure 1 used as the standard in this simulation study were dramatic. In practice, it will seldom be possible to find tests developed by a professional testing agency with parameter values that differ from the standard test in this study by a factor equal to two.

[Figure 3 about here]

The results in Figure 4 show that the presence of estimation error in the values of the parameters of the items did not have much impact, even for the condition with the larger errors. All curves were basically the same as those for the 40-item standard test in Figure 1, which were obtained without any estimation error in the values for the item parameters.

[Figure 4 about here]

The largest errors for the marginal equating method were obtained for the condition with an adaptive version of test Y in Figure 5. The reason is much more peaked conditional distributions of Y given θ for the adaptive version of the test than the version

with a fixed format. As a result, the marginal equating transformation become extremely sensitive to the population distribution of θ, which serves as the source of its bias. However, the results for the marginal transformation are presented only to make this effect visible, not to suggest any practical value. Real-life testing programs that use the marginal method to equate scores on an adaptive version of their test to a paper-and-pencil version typically use the estimated value of θ on the adaptive test and not the number-correct score for this purpose (Lawrence & Feigenbaum, 1997; Segall, 1997).

[Figure 5 about here]

Note that the marginal transformations in Figure 5 are defined over a smaller region of y than in the other conditions. This finding is also the result of the tendency of number-correct score on adaptive tests to be strongly peaked. As a result, the marginal transformations had to be truncated earlier at the value of .0001 for the joint probability function of X and Y.

As indicated before, the reason to look into the equating of an adaptive version of a test to a version with a fixed format was the fact that conditional methods seem much appropriate for this task. The results for the bias in the estimated conditional and posterior expected conditional transformations confirmed this expectation. The transformations had negligible bias over the region of values of y values for which the criterion of a probability of a score larger than .0001 was met. The application of these two methods in large-scale adaptive testing can be recommended without any hesitation.

Conclusion

The research in this paper was motivated by the fact that the error definition in the literature on observed-score equating allows only for sampling errors in the estimates of the distribution functions for the two versions of the test in the population of examinees. It was argued that, because the scores on one version of the test are transformed to be indistinguishable from the scores on another version, a definition of equating error based on the differences between these scores would be more natural. Embedding such

a definition in the framework of IRT led to the notion of conditional equating as well as to the formulation of two new methods of observed-score equating expected to perform better than traditional marginal equipercentile equating.

The results in the simulation studies were in agreement with this expectation. The conditional methods outperformed the marginal method under all conditions. Also, they never had a bias larger than one score point on the scale of the version of the test to which the scores were equated. The only exception was a condition with unrealistically low values for the discrimination parameter for the items in the test from which the scores were equated. On the other hand, marginal equipercentile equating yielded a bias in the equated scores that under some conditions reached a maximum as large as 10-15% of the maximum score on the test to which they were equated.

The results also showed that the conditional equating methods can be used to equate scores on an adaptive test to number-correct scores on a test with a fixed format, for example, a paper-and-pencil version of the test released for score-reporting purposes. Application of these methods in this context is natural because all items in the item pool are calibrated and it is a relatively simple step for the computer algorithm to produce a number-correct on the paper-and-pencil version of the test along with the ability estimate on the adaptive version.

References

Braun, H. I., & Holland, P. W. (1982). Observed score test equating: A mathematical analysis of some ETS equating procedures. In P. W. Holland & D. B. Rubin (Eds.). *Test equating* (pp. 9-49). New York: Academic Press.

Harris, D. B., & Crouse, J. D. (1993). A study of criteria used in equating. *Applied Measurement in Education, 6,* 195-240.

Kolen, M. J., & Brennan, R. L. (1995). *Test equating: Methods and practices.* New York: Springer-Verlag.

Lawrence, I, & Feigenbaum, M. (1997). *Linking scores for computer-adaptive and paper-and-pencil administrations of the SAT* (Research Report No, 97-12). Princeton, NJ: Educational Testing Service.

Liou, M., Cheng, P. E., & Li, M.-Y. (2001). Estimating comparable scores using surrogate variables. *Applied Psychological Measurement, 25,* 197-207.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Erlbaum.

Lord, F. M., & Wingersky, M. S. (1984). Comparison of IRT true-score and equipercentile observed-score "equatings". *Applied Psychological Measurement, 8,* 452-461.

Segall, D. O. (1997). Equating the CAT-ASVAB. IN W. A. Sands, B. K. Waters & J. R. McBride (Eds.). *Computerized adaptive testing: From inquiry to operation* (pp. 181-198). Washington, DC: American Psychological Association.

van der Linden, W. J. (2000). A test-theoretic approach to observed-score equating. *Psychometrika, 65,* 437-456.

van der Linden, W.J. (2001). Adaptive testing with equated number-correct scoring. *Applied Psychological Measurement, 25, 343-355.*

van der Linden, W.J., & Pashley, P. J. (2000). Item selection and ability estimation in adaptive testing. In W.J. van der Linden & C.A.W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 1-25). Norwell, MA: Kluwer Academic Publishers.

Wilks, M. B., & Gnanadesikan, R. (1968). Probability plotting methods for the analysis of data. *Biometrika, 55,* 1-17.

Wright, N. K., & Dorans, N. J. (1993). *Using the selection variable for matching or equating* (Research Report No. 92-3). Princeton, NJ: Educational Testing Service.

Zeng, L., & Kolen, M. J. (1995). An alternative approach for IRT observed-score equating of number-correct scores. *Applied Psychological Measurement, 19*, 231-240.

Figure Captions

Figure 1. Bias in (i) marginal equating transformations, (ii) estimated conditional equating transformations, and (iii) posterior expected conditional equating transformations at different values of θ for test lengths $n = 20$ (Panel a), $n = 40$ (Panel b) and $n = 60$ (Panel c).

Figure 2. Bias in (i) marginal equating transformations, (ii) estimated conditional equating transformations, and (iii) posterior expected conditional equating transformations at different values of θ for an easier (Panel a) and more difficult (Panel b) new version of the test, Y ($n = 40$).

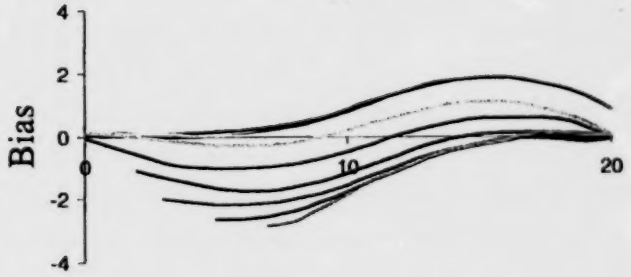
Figure 3. Bias in (i) marginal equating transformations, (ii) estimated conditional equating transformations, and (iii) posterior expected conditional equating transformations at different values of θ for a less discriminating (Panel a) and more discriminating (Panel b) new version of the test, Y ($n = 40$).

Figure 4. Bias in (i) marginal equating transformations, (ii) estimated conditional equating transformations, and (iii) posterior expected conditional equating transformations at different values of θ for smaller (Panel a) and larger (Panel b) estimation errors in the values of the item parameters for the two versions of the test ($n = 40$).

Figure 5. Bias in (i) marginal equating transformations, (ii) estimated conditional equating transformations, and (iii) posterior expected conditional equating transformations at different values of θ for an adaptive new version of the test, Y ($n = 40$).

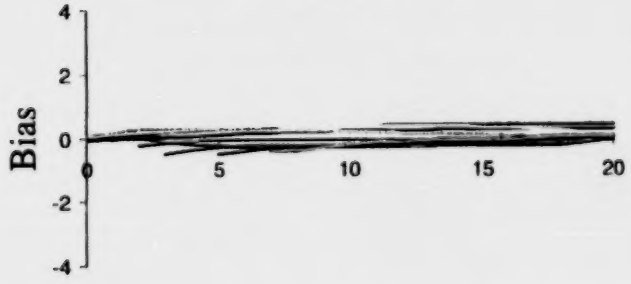
Panel a (n=20)

(i)



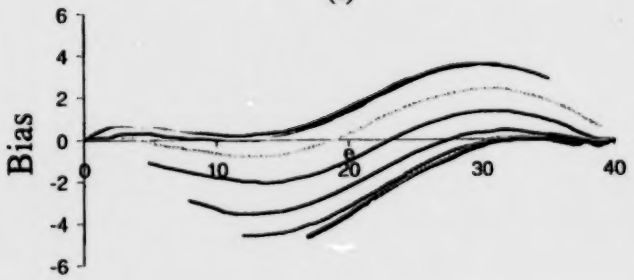
(ii)

(iii)

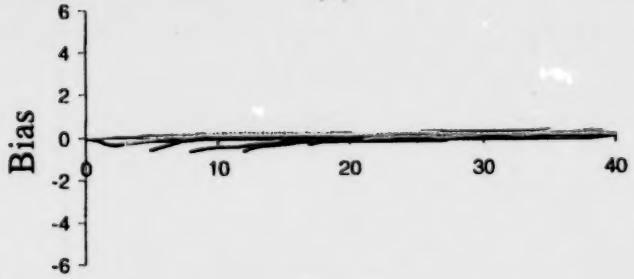


Panel b (n=40)

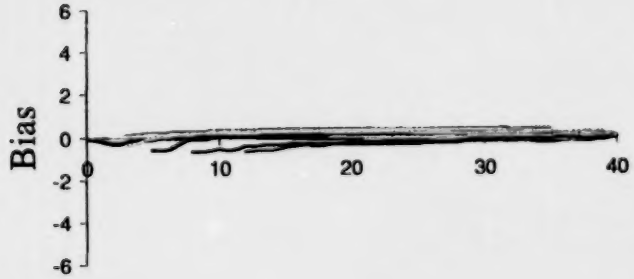
(i)



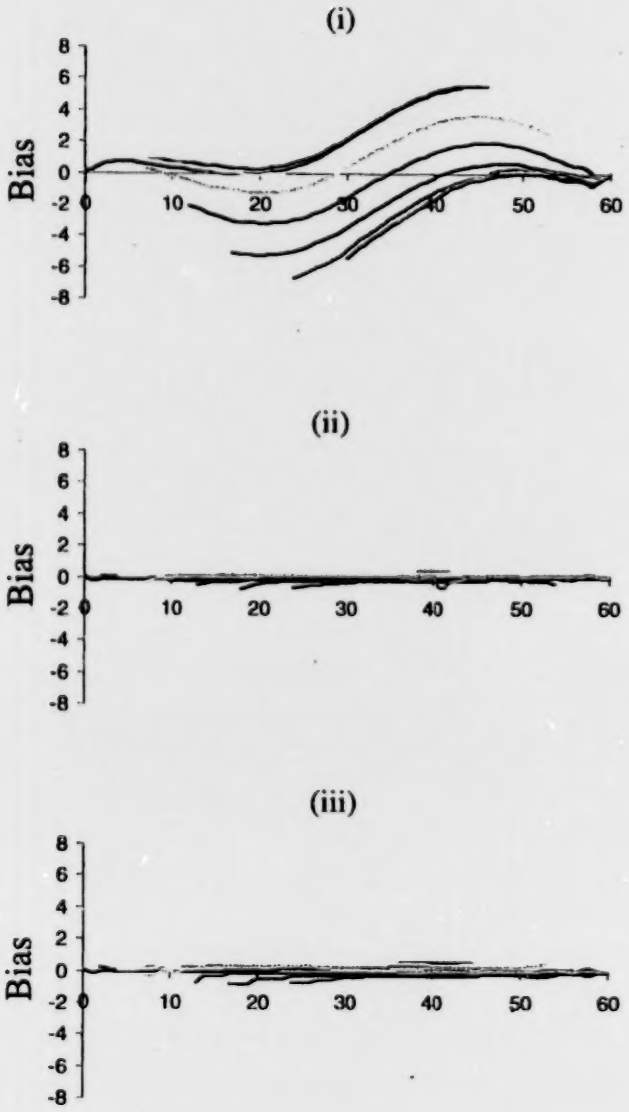
(ii)



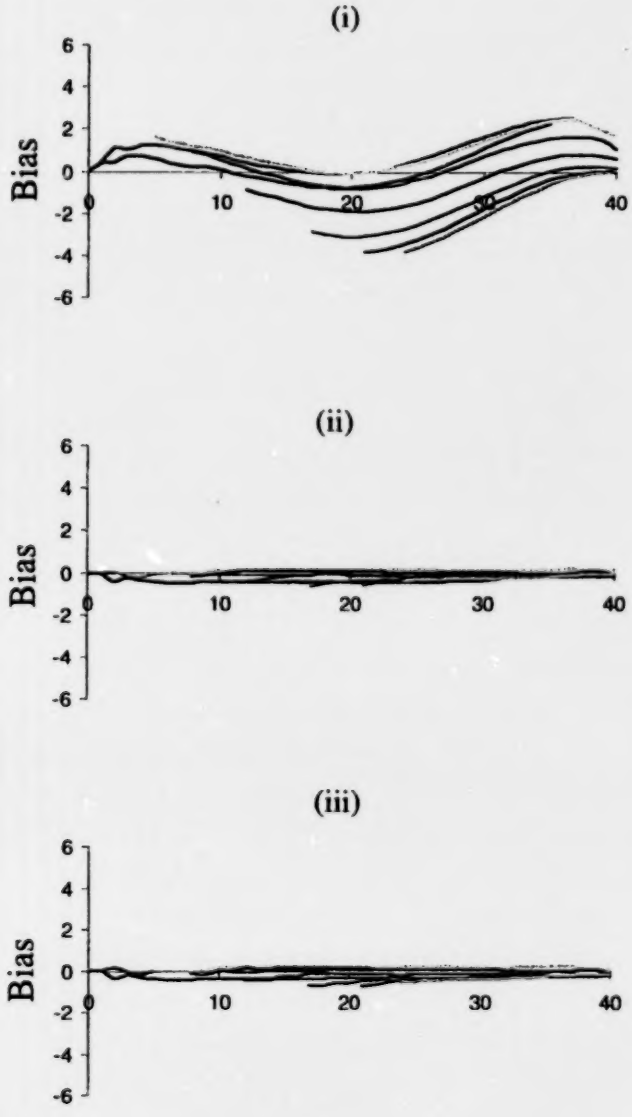
(iii)



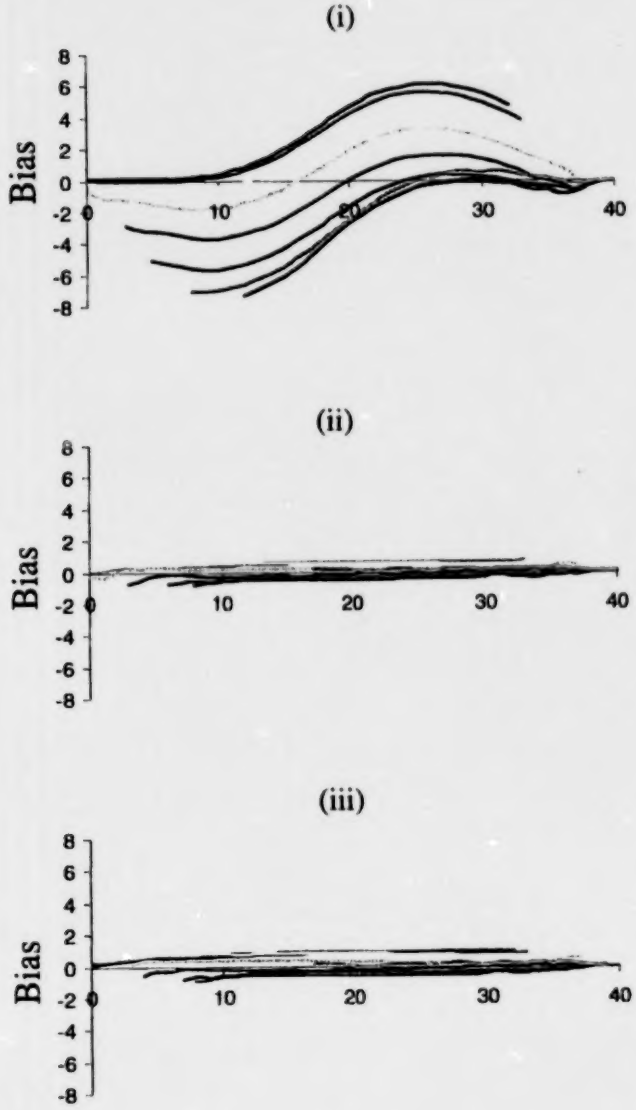
Panel c (n=60)


(i)
Bias
8
6
4
2
0
-2
-4
-6
-8
0
10
20
30
40
50
60
(ii)
Bias
8
6
4
2
0
-2
-4
-6
-8
0
10
20
30
40
50
60
(iii)
Bias
8
6
4
2
0
-2
-4
-6
-8
0
10
20
30
40
50
60

Panel a (version *Y* easier)

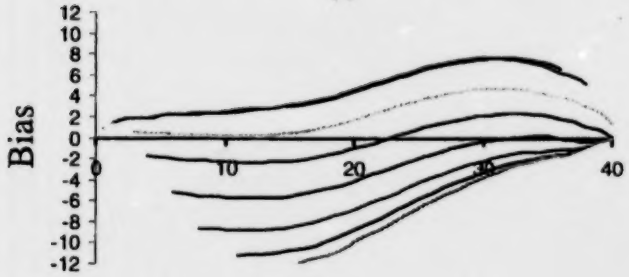


Panel b (version *Y* more difficult)

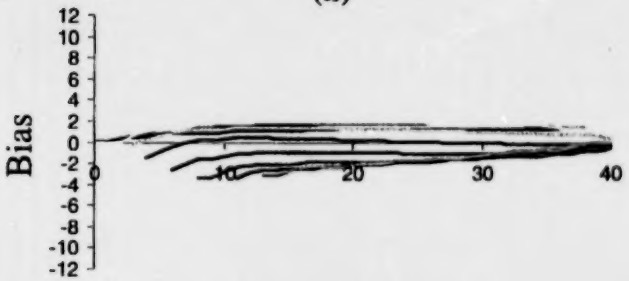


Panel a (version *Y* less discriminating)

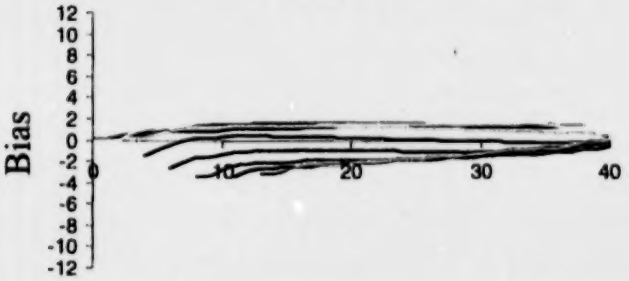
(i)



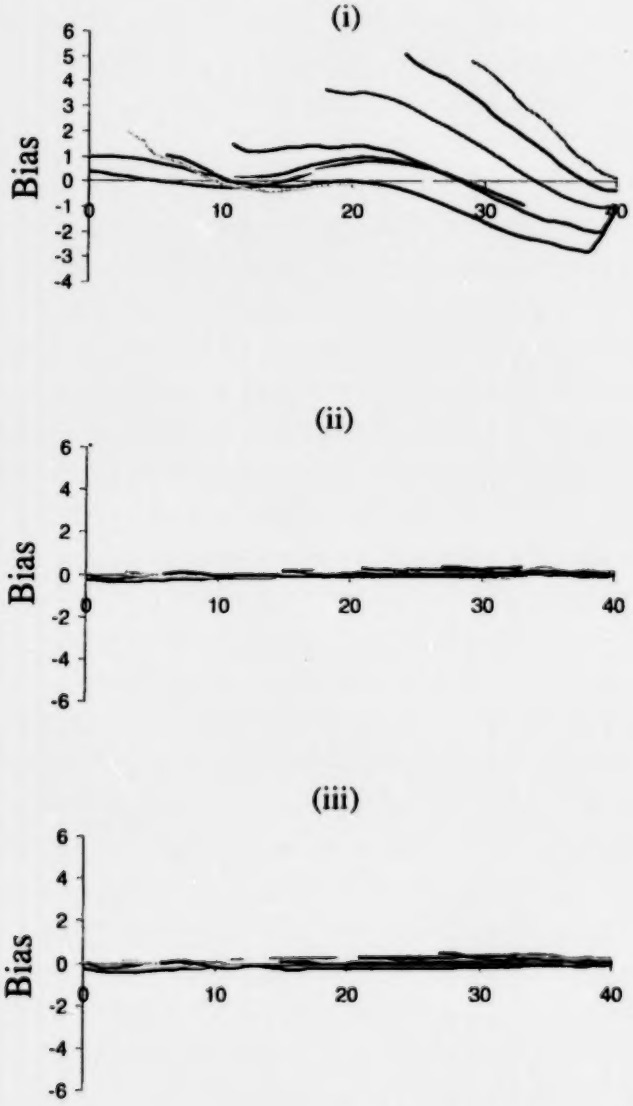
(ii)



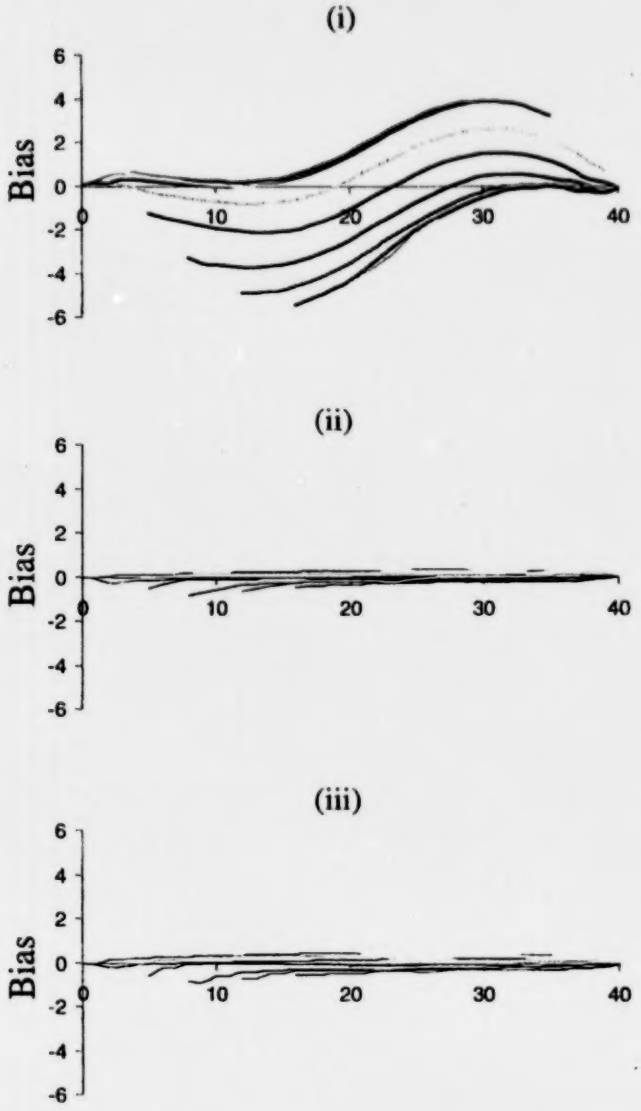
(iii)



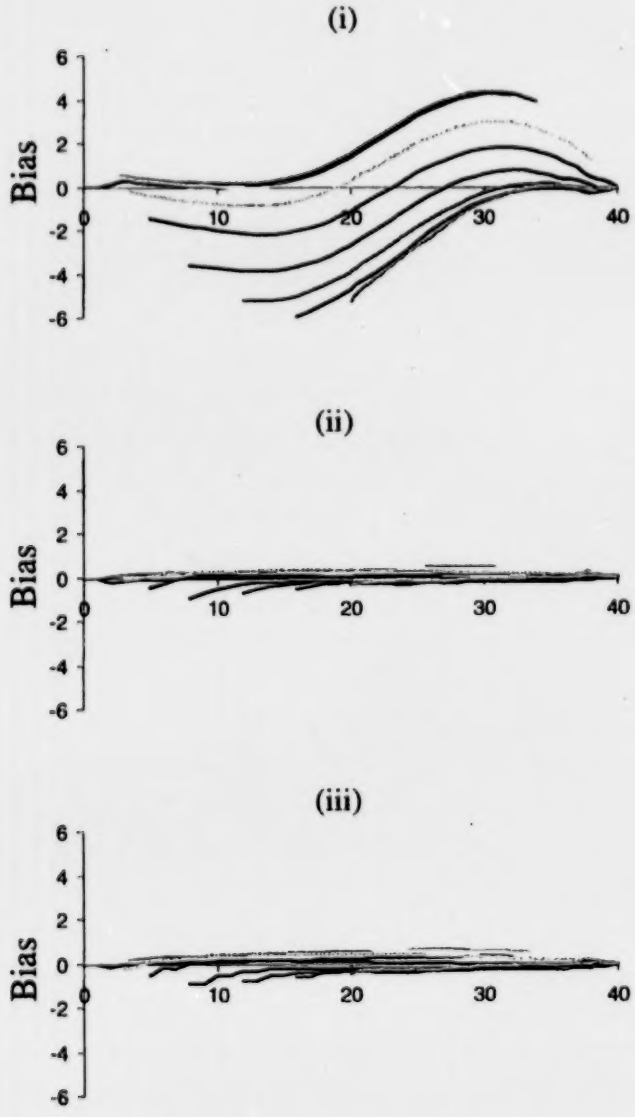
Panel b (version Y more discriminating)

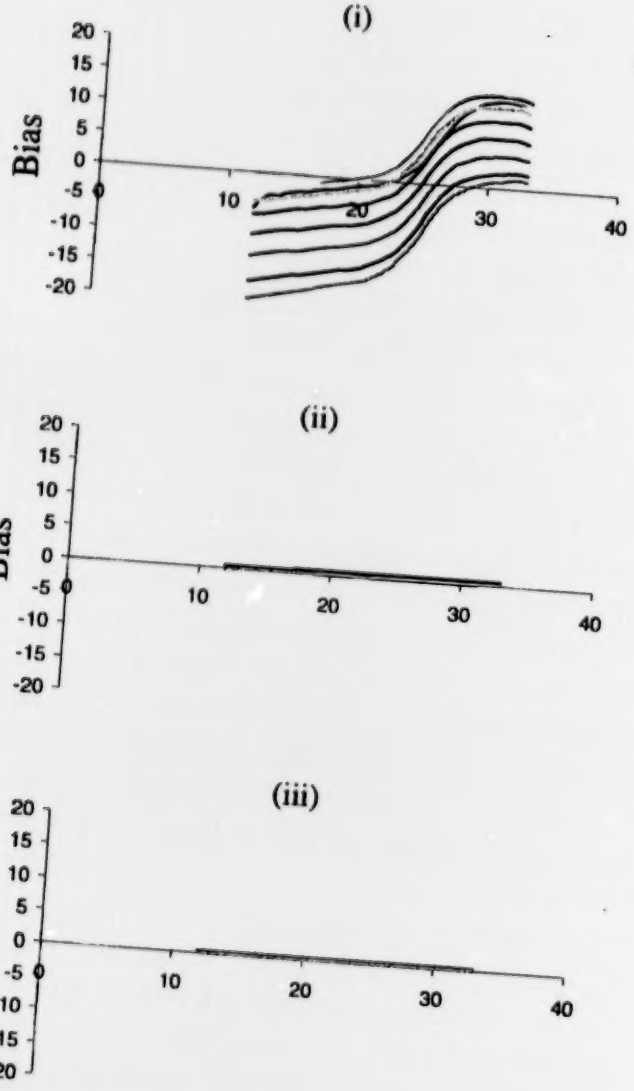


Panel a (both versions smaller error)



Panel b (both versions larger error)




(i)
Bias
20
15
10
5
0
-5
-10
-15
-20
10
20
30
40
(ii)
Bias
20
15
10
5
0
-5
-10
-15
-20
10
20
30
40
(iii)
Bias
20
15
10
5
0
-5
-10
-15
-20
10
20
30
40

Titles of Recent Research Reports from the Department of Educational Measurement and Data Analysis. University of Twente, Enschede, The Netherlands.

RR-02-03 W.J. van der Linden, *Estimating Equating Error in Observed-Score Equating*

RR-02-02 W.J. van der Linden, *Some Alternatives to Sympson-Hetter Item-Exposure Control in Computerized Adaptive Testing*

RR-02-01 W.J. van der Linden, H.J. Vos, & L. Chang, *Detecting Intrajudge Inconsistency in Standard Setting using Test Items with a Selected-Response Format*

RR-01-11 C.A.W. Glas & W.J. van der Linden, *Modeling Variability in Item Parameters in Item Response Models*

RR-01-10 C.A.W. Glas & W.J. van der Linden, *Computerized Adaptive Testing with Item Clones*

RR-01-09 C.A.W. Glas & R.R. Meijer, *A Bayesian Approach to Person Fit Analysis in Item Response Theory Models*

RR-01-08 W.J. van der Linden, *Computerized Test Construction*

RR-01-07 R.R. Meijer & L.S. Sotaridona, *Two New Statistics to Detect Answer Copying*

RR-01-06 R.R. Meijer & L.S. Sotaridona, *Statistical Properties of the K-index for Detecting Answer Copying*

RR-01-05 C.A.W. Glas, I. Hendrawan & R.R. Meijer, *The Effect of Person Misfit on Classification Decisions*

RR-01-04 R. Ben-Yashar, S. Nitzan & H.J. Vos, *Optimal Cutoff Points in Single and Multiple Tests for Psychological and Educational Decision Making*

RR-01-03 R.R. Meijer, *Outlier Detection in High-Stakes Certification Testing*

RR-01-02 R.R. Meijer, *Diagnosing Item Score Patterns using IRT Based Person-Fit Statistics*

RR-01-01 H. Chang & W.J. van der Linden, *Implementing Content Constraints in Alpha-Stratified Adaptive Testing Using a Shadow Test Approach*

RR-00-11 B.P. Veldkamp & W.J. van der Linden, *Multidimensional Adaptive Testing with Constraints on Test Content*

RR-00-10 W.J. van der Linden, *A Test-Theoretic Approach to Observed-Score Equating*

RR-00-09 W.J. van der Linden & E.M.L.A. van Krimpen-Stoop, *Using Response Times to Detect Aberrant Responses in Computerized Adaptive Testing*

RR-00-08 L. Chang & W.J. van der Linden & H.J. Vos, *A New Test-Centered Standard-Setting Method Based on Interdependent Evaluation of Item Alternatives*

RR-00-07 W.J. van der linden, *Optimal Stratification of Item Pools in a-Stratified Computerized Adaptive Testing*

RR-00-06 C.A.W. Glas & H.J. Vos, *Adaptive Mastery Testing Using a Multidimensional IRT Model and Bayesian Sequential Decision Theory*

RR-00-05 B.P. Veldkamp, *Modifications of the Branch-and-Bound Algorithm for Application in Constrained Adaptive Testing*

RR-00-04 B.P. Veldkamp, *Constrained Multidimensional Test Assembly*

RR-00-03 J.P. Fox & C.A.W. Glas, *Bayesian Modeling of Measurement Error in Predictor Variables using Item Response Theory*

RR-00-02 J.P. Fox, *Stochastic EM for Estimating the Parameters of a Multilevel IRT Model*

RR-00-01 E.M.L.A. van Krimpen-Stoop & R.R. Meijer, *Detection of Person Misfit in Computerized Adaptive Tests with Polytomous Items*

RR-99-08 W.J. van der Linden & J.E. Carlson, *Calculating Balanced Incomplete Block Designs for Educational Assessments*

RR-99-07 N.D. Verhelst & F. Kaftandjieva, *A Rational Method to Determine Cutoff Scores*

RR-99-06 G. van Engelenburg, *Statistical Analysis for the Solomon Four-Group Design*

RR-99-05 E.M.L.A. van Krimpen-Stoop & R.R. Meijer, *CUSUM-Based Person-Fit Statistics for Adaptive Testing*

RR-99-04 H.J. Vos, *A Minimax Procedure in the Context of Sequential Mastery Testing*

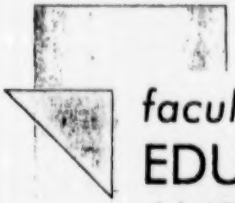
RR-99-03 B.P. Veldkamp & W.J. van der Linden, *Designing Item Pools for Computerized Adaptive Testing*

RR-99-02 W.J. van der Linden, *Adaptive Testing with Equated Number-Correct Scoring*

RR-99-01 R.R. Meijer & K. Sijtsma, *A Review of Methods for Evaluating the Fit of Item Score Patterns on a Test*

RR-98-16 J.P. Fox & C.A.W. Glas, *Multi-level IRT with Measurement Error in the Predictor Variables*

Research Reports can be obtained at costs, Faculty of Educational Science and Technology, University of Twente, TO/OMD, P.O. Box 217, 7500 AE Enschede, The Netherlands.



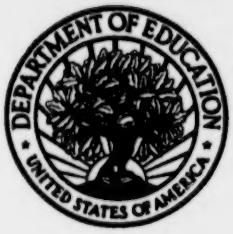
faculty of

EDUCATIONAL SCIENCE AND TECHNOLOGY

A publication by
The Faculty of Educational Science and Technology of the University of Twente
P.O. Box 217
7500 AE Enschede

BEST COPY AVAILABLE

TM034738



U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

NOTICE

Reproduction Basis

This document is covered by a signed "Reproduction Release (Blanket)" form (on file within the ERIC system), encompassing all or classes of documents from its source organization and, therefore, does not require a "Specific Document" Release form.

This document is Federally-funded, or carries its own permission to reproduce, or is otherwise in the public domain and, therefore, may be reproduced by ERIC without a signed Reproduction Release form (either "Specific Document" or "Blanket").